Bridging the gap between culture studies and computational linguistics

Inge van de Ven[1][0000−0001−7237−3779] and
Menno van Zaanen[2][0000−0003−1841−2444]

[1] Tilburg University, Tilburg, The Netherlands
i.g.m.vdven@tilburguniversity.edu
[2] South African Centre for Digital Language Resources, Potchefstroom, South Africa
menno.vanzaanen@nwu.ac.za

Abstract. Since 2015, the Tilburg School of Humanities and Digital Sciences at Tilburg University has provided yearly funding for two student assistants in a Research Traineeship Program. The aim of this program is to provide the student assistants with practical research experience by allowing them to work together with two experienced researchers in a practical research project. The researchers who apply for funding in this program have to work at different departments within the school, effectively encouraging cross-departmental collaborations. Within this interdisciplinary research project, the student assistants build practical research experience by following the full research process from developing research questions to presenting and publishing results. Also, due to the interdisciplinary nature, the student assistants (as well as the supervising researchers) are confronted with different research methodologies and cross-disciplinary problems. Here, we describe one of the research traineeship projects, the problem tackled, approaches used, and results. The project proposal contained ideas on how to bridge the gap between close (focused on detail, small amounts of text) and distant reading (focused on the larger picture, large amounts of text). However, specific research questions still needed to be designed. Research questions and corresponding methodologies that combine the fields of culture studies and computational linguistics were selected after extensive discussions. As the discussions demonstrated points of terminological ambivalence, we decided to subdivide the project along disciplinary lines. This resulted in a culture studies and a computational linguistics sub-project; these were constantly aligned during the project. This led to new insights into both the culture studies and computational linguistics methodologies. Here, we focus on the problems that were encountered during the project, which can mostly be attributed to difficulties related to (discipline-specific) terminology. Based on these experiences, we also provide suggestions for future work and recommendations for similar interdisciplinary collaborations.

Keywords: Qualitative linguistics · Computational linguistics · Digital Humanities · Close reading · Distant reading

1 Introduction

Research in the field of Digital Humanities is often performed as collaborative projects with researchers from various fields. This interdisciplinary approach can lead to very interesting results, but can also pose specific challenges. In this article, we describe our experiences with interdisciplinary research in the field of digital humanities (with researchers from the fields of culture studies and computational linguistics), where the project team also included two student assistants (one from each field).

We first provide a brief overview of the context in which this research took place, describing the research aims, and specific funding that was received. Next, we describe of the development of the research questions that were tackled in this project. The interdisciplinary nature of the research had an effect on the approach and methodologies used and we discuss the design choices made. Next, we describe the results of the research and the effect on the different fields. Even though we describe the actual research, this article focuses mostly on the problems and solutions related to the collaboration between the researchers. Hence, we provide a more extensive description of the experience of the collaboration. Finally, we provide a short conclusion and recommendations for both educators as well as other researchers who are at the start of similar situations.

2 Context

Starting 2015, the School of Humanities and Digital Sciences at Tilburg University (then called School of Humanities) supported the so called Research Traineeships Program. This program was initially called the KNAW Academy Assistants Program, which was funded by the KNAW (Royal Netherlands Academy of Science). Over the years (including the 2020–2021 academic year), this program has funded 49 research projects.

This funding program was specifically designed for researchers working at the School of Humanities and Digital Sciences and provides funding for two student assistants to work together with two experienced researchers from different departments for the full academic year. The program had several aims. First, it provides the student assistants with insight and experience in daily research practice. The student assistants are expected to perform research together with the two experienced researchers in a team. Second, the student assistants are expected to present their findings at a symposium at the end of the year, providing experience in presenting. Finally, the research should work towards a collaborative academic publication at the end of the project.

Both authors collaborated as supervisors twice in this program. Van de Ven works at the Department of Culture studies and has a background in literary

studies with a specific interest in different modes of reading. Van Zaanen at that time worked at the Department of Communication and Information Sciences (which was later renamed as the Department of Cognitive Science and Artificial Intelligence) and performed research in the field of (applied) computational linguistics. In the academic year 2017–2018, the authors received funding for a project called "Negotiating legibility: Bridging the gap between close and distant reading". In the academic year 2018–2019 this research was extended within a second funded project called "Using human in the loop to bridge the gap between close and distant reading". Here, we discuss the experiences related to the first project.

Once the project proposal was approved, the authors searched for student assistants for the project. These students should study within the School of Humanities and Digital Sciences, but no restrictions on the educational level were set (bachelor, pre-master, master). Different channels were used to contact students, including emails, sharing information at lectures, etc. In the end, two students were hired: Christoph Aurnhammer, who was a research master student language and communication (which is a combined program between Tilburg University and Radboud University, Nijmegen, the Netherlands) with an interest in computational approaches and Iris Cuppen, who was a master student culture studies. The students provided a similar balance between fields of study as the researchers.

3 Research problem

The project we describe in this article takes places in a larger project that the two authors embarked on. This larger project deals with a central issue in textual analysis within the context of Digital Humanities (DH).[3]

For years, the uses and misuses of DH have been fiercely debated. In 2016 Daniel Allington, Sarah Brouillette and David Golumbia published a much-discussed piece [1], critiquing Digital Humanities advocates for their alignment with the "neoliberal takeover" of universities, provoking Alan Liu to defend DH on Twitter.[4]

A recurring point of polarization in this debate is the valuation of "distant reading" versus the defense of "close reading". Since 2000, many have followed Franco Moretti's provocative call for distant reading. Moretti deemed close reading "a theological exercise" and urged humanists to "read less" [7, p. 48]. Others, like Michael Manderino [5] and Antoine Compagnon [4], attempt to rehabilitate close reading, arguing that we need its associated skills and strategies more than ever in our media-saturated age.

The issue is to a considerable extent a crisis of attention. Close reading is a strategy that entails devoting attention to minute details. Some have argued that

[3] This and related research has been published in more detail [2, 9, 10]. Here we provide a description of this research in the context of our experiences regarding the collaboration.

[4] See https://storify.com/ayliu/on-digital-humanities-and-critique.

such strategies stand to be regained and revalued in a time when we encounter vast bodies of information through multiple platforms. Yet, as Moretti has noted [7, p. 57], close reading necessarily implies a select canon, or a small slice out of the available data. Acts of selection are losing currency as big data theorists today deem sampling "an artifact of a period of information scarcity, a product of the natural constraints on interacting with information in an analog era" [6, p. 16–17], and as companies like Google strive to collect and organize the world's information [8, p. 2]. A close-up perspective pertains to the small; distance allows us to see the bigger picture, and the latter is currently privileged.

However, is a defense of the old humanist strategy of close reading the only way out of this polemic? Are quantified, big-scale methodologies and meticulously attentive readings mutually exclusive? If not, how do we bridge the gap and unite the most valuable properties of both approaches to texts? Our larger, ongoing interdisciplinary research project seeks to reflect on and develop analytical instruments that combine classical-humanist attention to the singular object with methods applicable to variable scales of textuality.

How do we engage with literary/media objects (either minimalist or maximalist, digital or analog) that resist this binary between close or distant reading, and demand a variation between scales? Qualitative, traditional humanities methods of textual analysis fall short of analyzing thousand- or million-page works, micro narratives, or Twitterbot poetry: such objects demand new ways of reading. In contrast, quantitative DH methods like distant and algorithmic reading (see [3, 7]) often tend to undervalue specific features in favor of larger trends and patterns.

In our ongoing collaborations, we aim to analyze corpora of literary/textual objects from the minimalist to the maximalist that solicit readings which zoom in and out between part and whole, micro and macro, surface and depth. On the basis of these, we propose reading strategies that move beyond the dichotomy and allow us to oscillate between the close and the distant, small and large-scale, minimalist and the maximalist, deep and hyper attention. This allows us to add grey scales to the originally black/white distinction between close and distant reading.

This investigation requires the development of computational tools that deal with large amounts of documents (e.g., topic identification, automatic summarization). However, in order to evaluate the computational analysis, in-depth, qualitative analyses of the performance of the computational analyses are essential. Effectively, this combines close and distant reading, enabling close reading of "salient" documents that either describe common trends in large document collections or provide views that counter this common trend.

3.1 Selection of specific problem

Given the overall project description provided in the previous section, a smaller sub-project needed to be defined for the research of the Research Traineeship Program. This sub-project should allow the student assistants (one with a computational background and one with a culture studies background) to conduct

research in their respective fields while engaging in dialogue with, and learning from, the other field. As the project description provided a rather general direction, specific methodological choices were still required. Intensive discussions with the researchers and student assistants took place at the beginning of the project.

Main difficulties at this stage related to unclear definitions and expectations from the different research backgrounds, for instance, regarding terms as "close reading" and "distant reading". This became more apparent during the project. For instance, the computational linguistics researchers linked distant reading with automatic summarization, whereas to the culture studies researchers this related to a more abstract approach as well as a provocation and a research agenda in literary studies.

To resolve this issue, we decided on topics that could be split into two approaches, requiring their own methodologies, and split the team based on research background of the researchers. The computational linguistics researchers focused on methodologies related to quantitative aspects whereas the culture studies researchers dealt with methodologies related to qualitative aspects. Regular meetings allowed for the continuous alignment between the two research areas. Throughout the project, these meetings also helped in better understanding the terminology and methodologies of the different fields.

3.2 Specific problem description

The main methodological research problem that this project sought to tackle was how to bridge the gap between close and distant reading. A brief survey of how different scales of analyses are connected in actual literary-critical practice showed us that close and distant reading were never mutually exclusive. This is part of our overall research aim to find ways to mix qualitative and quantitative methods and use them to analyze corpora that solicit readings that zoom in and out between part and whole.

As a practical problem, we envisioned a computational approach that might help or even replace manual annotation of texts on a pragmatic level. Whereas humans can analyze the function of a particular text, doing this for huge amounts of text is simply impractical. As computational approaches do not get tired, applying such techniques may allow for a similar analysis where human analysis is not practical. However, for this to work, the computational approaches should lead to similar results as the manual, human analysis.

4 Approach

We developed a strategy for using manual annotation to evaluate and supplement a Latent Dirichlet Allocation (LDA) model. Rather than following a top-down approach, in our mixed method, manual annotation is not of a sample that follows from the overview produced by the distant reading as is often the case. Our analysis incorporated local annotation in a distant reading; if the computational

technique behaves similarly to the manual approach, we know we can effectively use a distant reading technique to complement the close reading analysis. Here, we briefly outline our approach. For a more extensive step-by-step report, see [2].

We focused on a discussion thread of the online forum Reddit, from January 19$^{\text{th}}$ of 2017, which posed the questions: "Should the Democrats nominate a celebrity in 2020? What would be the pros and cons?" Our data set consists of 449 (461 including deleted comments) responses to these questions. We employed a hypothesis-free form of close reading: we first started looking for patterns in the material of the discussion in a rather open-ended way, without explicitly framing our expectations. This process took place in a bottom-up fashion: two human annotators analyzed the posts in the thread, and chose a word or phrase to summarize each post, developing a collection of labels. Based on this, fifteen classes were identified: ten related to three different underlying questions that we formulated based on the classes. After manually annotating the posts separately, we compared the lists of the outcome and slightly modified the categories to accommodate both our findings. There was a high degree of congruence between the categories assigned in both annotations.

Separately, we adopted a data-driven ideal of distant reading. The approach was unsupervised, and as Reddit threads are open-ended, the number of topics in the shape of clusters resulting from the distant reading algorithm needed to be variable. We chose LDA topic modeling as it allows for automatic grouping of documents according to latent content categories. Because it is unknown how many topics are to be expected from a discussion thread, a pass through a range of numbers of LDA topics was necessary. Investigating a whole range of numbers of topics may additionally reveal several different levels of topical granularity that can be captured using LDA. Each post in the thread was treated as a separate document. For each, the LDA topic with the highest probability was selected. This was a limiting decision, yet we found that the per-document probability distributions usually strongly favor one single topic with a very high probability, while the probabilities for the other topics are close to zero.

Documents were thus represented by two labels: a close reading annotation and a (single) LDA topic. We aimed to identify LDA topics and manual annotations that express similar concepts. Going through the list of documents, co-occurrences of the two annotations were counted in a matrix, with the aim to find, for each LDA topic, the manual annotation class that is best represented. During normalisation, the counts were divided by the class support of the manual annotation class of this row, resulting in fractions instead of absolute numbers.

For each pair of classes with highest co-occurrence, we generated a mapping from LDA classes to manual annotation classes. In the original list of documents, each document was represented by two annotations that are drawn from one common pool of possible annotations. To measure the extent to which the predictions and the gold standard overlap, we calculated the accuracy. We produced topic models with a number of topics ranging from 1 to N, with N equal to the number of documents in the collection. Overlap accuracy was calculated

for models with 1 to 461 topics. The accuracy increased with higher numbers of topics in the LDA models. However, the model with the highest overlap accuracy has low generalizing power. It moves away from the desired distant reading perspective, which would group the comments into a limited set of categories. We looked at a combination of the forward and reverse accuracy and balance the two LDA to manual assignment and manual to LDA assignment scores and inspect where the two curves intersect: the number of topics with the lowest absolute difference in accuracy. As baseline we used a random assignment of manual annotations to documents and compare them to the actual manual assignment. This process was bootstrapped and averaged over 1000 trials. Regardless of whether the co-occurrence matrix is normalised for class support the LDA classes lead to a higher overlap with the manual annotations than the random assignment. The overlap between the LDA classes and the manual annotations is thus above chance level.

5 Results

Based on the close reading strategy described above, we identified fifteen classes of posts. While the initial question of the Reddit thread regards viewpoints on the pros and cons of a future celebrity president and the names of potential Democrat candidates, the thread soon developed into a more complex conversation in which different "new" questions were discussed. The manual annotation revealed that we did not only identify several opinions or find information that regards the question that the thread set out from; in addition, we found another class of posts that we choose to describe along the lines of discourse function: for instance, humor, hyperlinks, or direct responses (e.g., "yes", "no", "indeed"). These are elements that fulfill a certain function within the larger collective discourse, without being reducible to an answer to the main question, an opinion or a piece of information. The posts that fall within the function classes do not correspond to specific topics, so it is likely that these elements can only be identified by close reading. In particular, posts that deal with humor and irony are hard to identify computationally, which underlines the importance of human annotation. The bottom-up process of manual annotation and our hypothesis-free form of close reading unraveled underlying questions and contexts that we can use to aid computational strategy.

There are many questions on the close reading side, leading to different collections of close reading classes and we need to determine which of those correspond to the identified distant reading classes. In other words, it is as of yet unclear how generalizable this approach is. We have proposed a way to find a mapping from LDA to manually assigned classes. The analysis of the close and distant reading demonstrates that LDA is indeed a possibility to generate a computational model of close reading annotations. However it needs to be clearly stated that the overlap between the two sets of labels, as measured by accuracy, is relatively low. This may, first, be due to the rather high number of 15 classes of comments on a relatively small data set (461 comments). Second, some of

the close reading based classes describe function rather than content. LDA topic modeling is designed to primarily capture content information and may thus not be able to accurately capture functional classes. Still, our approach led to an overlap above chance level that could be increased.

6 Collaboration

The collaboration between the researchers involved in this project can roughly be divided into three phases. In the first phase, the two authors discussed the possibility of writing a project proposal. During the second phase, all four researchers discussed the details of the research to be performed and the writing of publications represents the third phase.

During the first phase, the two authors met several times. Van de Ven proposed a topic from the field of culture studies, which Van Zaanen tried to match up with techniques from the field of computational linguistics. Both researchers had an approximate idea of the other researcher's fields, but they were no experts of the other's field of expertise. During these discussions, several topics were discussed, which included the researchers trying to explain to each other the challenges and possibilities of the problems, potential research direction, and aims. When discussing a topic, which typically started from a problem in the field of culture studies, a brief explanation of the problem and related terminology was given. Based on this description, potential computational techniques were mentioned and a brief explanation of the possibilities and limitations of the techniques was provided. In the end, a relatively broad topic was selected, which, on the one hand, was thought to to be concrete enough as the basis for research, but, on the other hand, general enough so that specific methodological choices could still be made (also based on the interest of the student assistants).

The second phase started directly after the student assistants joined the project. As the students also had their preferences and expertise, we expected with their influence the topic to become more detailed quickly. These discussions made it clear that the terminology from both the fields of culture studies and computational linguistics are quite far apart. It is difficult to explain the exact details of the terminology without having a more complete background in that field. For instance, it was difficult to define the terms close and distant reading or even to explain these concepts to computational linguists. Similarly, it turned out to be difficult to quantify the possibilities and limitation of techniques such as automatic summarization (which the computational linguists initially thought equated to distant reading) or to explain the possibilities and limitations of LDA without going into too much detail.

To circumvent the problems of trying to explain the different terms and techniques in too much detail, we decided to select a particular problem and split the approach in two: a culture studies approach and a computational linguistics approach. This allowed us to use the techniques we were familiar with, without having to rely on the other researchers to fully understand at that point what the techniques does or how it should be applied. At the same time, we organized

regular meetings, sharing the results and experiences with each other. These meetings were very useful as the intermediate results allowed for an alignment of the approaches for research purposes, but they also provided examples of the terms that were unclear earlier. For instance, the close reading, manual annotation of the Reddit posts provided insight in the culture studies approach, whereas initial results of the LDA system also indicated expected results from the computational linguistics techniques. This allowed us to think about how to compare and evaluate the different techniques, bringing together both culture studies and computational linguistics results.

Once the results of both approaches were available and compared against each other, all researchers participated in writing publications and preparing presentations [2, 10]. The culture studies and computational linguistics researchers each wrote initial descriptions of their approach and the rest was written collaboratively. By this time, the different researchers understood enough of the entire topic, corresponding terminology and techniques to be able to rewrite and improve on the entire text.

To summarize the process, the authors initially expected that it would be easier to collaborate. Having experience in culture studies (e.g., in close and distant reading, big data) and computational linguistics (e.g., text analysis) respectively, we assumed the boundaries between our research fields would not be too far apart. However, it turned out that it is relatively easy to get an approximate idea of the concepts in the other's research field, but doing research requires more detailed knowledge. Separating the approaches allowed the researchers from both fields to work within their own area of expertise. At the same time, however, intermediate results allowed for the alignment of topics in the different fields. In general, this collaboration has provided new insights into how research is performed in different research areas. Not only did the student assistants receive training on research in their own research field, they experienced research using methodologies from other fields. The same also holds for the supervisors, who are still working together.

7 Conclusions and recommendations

In this article, we described a highly interdisciplinary research project. The project was funded through the Research Traineeship Program, which required cross-departmental collaboration. The research proposal, which formed the basis for the research described here, was written by researchers from the fields of culture studies and computational linguistics and provided funding for two student assistants to collaborate within the interdisciplinary research project. The student assistants were expected to function as collaborators within the project, providing them with practical research experience.

The main challenge in this interdisciplinary research project turned out to be related to discipline-specific terminology. It is very difficult to properly define many of the terms used in a particular research field, even to researchers who come from a closely related field. Typically, more background knowledge is

required to properly appreciate and understand the exact meaning of the terms from a particular research area.

During the project, the research has been split into sub-projects related to the different backgrounds of the researchers. This allowed for the researchers to do research with known techniques. Intermediate results were shared with all project members, providing more insight in the (im)possibilities of the techniques, but also in the meaning of the terminology that was initially unclear.

For future research, we encourage researchers to cross the boundaries of their own research area. Interdisciplinary research allows for the investigation that transcends individual research areas and may lead to very interesting research results. In particular, working together with student assistants (allowing them to participate as researchers) brings in fresh views. As much of the research area is still new to them, they often ask questions that highlight interesting open problems.

Experience from this interdisciplinary research has shown that care has to be taken regarding terminology. It is often difficult to define terms that are considered fundamental or clear in a particular research area, if the required background knowledge is missing. In the project described here, we resolved this issue by splitting the project into discipline-specific sub-projects, while the alignment of the results during the project helped in providing a better understanding of the terminology.

References

1. Allington, D., Brouillette, S., Golumbia, D.: Neoliberal tools (and archives): A political history of digital humanities. LA Review of Books **1** (2016)
2. Aurnhammer, C., Cuppen, I., van de Ven, I., van Zaanen, M.: Manual annotation of unsupervised models: Close and distant reading of politics on reddit. DHQ: Digital Humanities Quarterly **13**(3) (2019)
3. Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., Schnapp, J.: Digital_Humanities. Mit Press (2012)
4. Compagnon, A.: The resistance to interpretation. New Literary History **45**(2), 271–280 (2014)
5. Manderino, M.: Reading and understanding in the digital age. a look at the critical need for close reading of digital and multimodal texts. Reading Today pp. 22–23 (Jan/Feb 2015)
6. Mayer-Schönberger, V., Cukier, K.: Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt, New York (2013)
7. Moretti, F.: Distant Reading. Verso, London (2013)
8. Vaidhyanathan, S.: The Googlization of everything:(and why we should worry). University of California Press, Berkeley and Los Angeles (2012)
9. van de Ven, I., Lim, C., Steenbakker, M., van Zaanen, M.: Negotiating close and distant reading: Heteroglossia and networks in Zadie Smith's White Teeth. In: Digital Humanities Benelux (2018)
10. van de Ven, I., van Zaanen, M., Cuppen, I., Aurnhammer, C.: Analyzing reading strategies: bridging the gap between close and distant reading. In: DH Benelux. Utrecht, the Netherlands (Jul 2017)